Speech Communication 127 (2021) 1–16

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic quality control and enhancement for voice-based remote Parkinson's disease detection

Amir Hossein Poorjam [a,*], Mathew Shaji Kavalekalam [a], Liming Shi [a], Jordan P. Raykov [b], Jesper Rindom Jensen [a,1], Max A. Little [c,d], Mads Græsbøll Christensen [a,1]

[a] Audio Analysis Lab, CREATE, Aalborg University, Aalborg 9000, Denmark
[b] School of Engineering and Applied Science, Aston University, Birmingham, UK
[c] School of Computer Science, University of Birmingham, Birmingham, UK
[d] Media Lab, MIT, Cambridge, Massachusetts, USA

ARTICLE INFO

Keywords:
Acoustic mismatch
Parkinson's disease detection
Quality control
Speech enhancement

ABSTRACT

The performance of voice-based Parkinson's disease (PD) detection systems degrades when there is an acoustic mismatch between training and operating conditions caused mainly by degradation in test signals. In this paper, we address this mismatch by considering three types of degradation commonly encountered in remote voice analysis, namely background noise, reverberation and nonlinear distortion, and investigate how these degradations influence the performance of a PD detection system. Given that the specific degradation is known, we explore the effectiveness of a variety of enhancement algorithms in compensating this mismatch and improving the PD detection accuracy. Then, we propose two approaches to automatically control the quality of recordings by identifying the presence and type of short-term and long-term degradations and protocol violations in voice signals. Finally, we experiment with using the proposed quality control methods to inform the choice of enhancement algorithm. Experimental results using the voice recordings of the mPower mobile PD data set under different degradation conditions show the effectiveness of the quality control approaches in selecting an appropriate enhancement method and, consequently, in improving the PD detection accuracy. This study is a step towards the development of a remote PD detection system capable of operating in unseen acoustic environments.

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disorder which progressively makes the patient unable to control the movement normally and, consequently, decreases the patient's quality of life (Ishihara et al., 2007). Although there is no cure to stop the process of neurodegenerative progression in a PD patient, medications, in early stages, and surgeries, in advanced stages of the disease, can decelerate the PD progression and improve the patient's functional capacity (Singh et al., 2007). Voice and speech problems, such as soft voice, monotonous pitch, hoarse voice quality, change in rate of speech, and imprecise articulation are typically the first appearing symptoms in PD patients (Murdoch, 1998). It has been demonstrated in Eliasova et al. (2013) that acoustic analysis of voice signals can better reflect small changes in PD progression than perceptual evaluation of voice by a therapist. This has motivated researchers to take advantage of advanced speech signal processing and machine learning algorithms to develop highly accurate and data-driven methods for detecting PD symptoms from voice signals (Tsanas et al., 2012; Zhan et al., 2016; Gil and Johnson, 2009). Moreover, advances in smart phone technology provide new opportunities for remote monitoring of PD symptoms by bypassing the logistical and practical limitations of recording voice samples in controlled experimental conditions in clinics (Rusz et al., 2018; Zhan et al., 2016). However, there is a higher risk outside controlled lab conditions that participants may not adhere to the test protocols, which probe for specific symptoms, due to lack of training, misinterpretation of the test protocol or negligence. Moreover, voice signals in remote voice analysis might be subject to a variety of degradations during recording or transmission. Processing the degraded recordings or those which do not comply with the assumptions of the test protocol can produce misleading, non-replicable and non-reproducible results

* Corresponding author.
E-mail addresses: ahp@create.aau.dk (A.H. Poorjam), msk@create.aau.dk (M.S. Kavalekalam), ls@create.aau.dk (L. Shi), y.raykov@aston.ac.uk (J.P. Raykov), jrj@create.aau.dk (J.R. Jensen), maxl@mit.edu (M.A. Little), msg@create.aau.dk (M.G. Christensen).

[1] EURASIP member.

[2] In this paper, by "signal enhancement", we refer to all algorithms intended to enhance the quality of degraded signals.

https://doi.org/10.1016/j.specom.2020.12.007
Received 16 August 2020; Received in revised form 9 November 2020; Accepted 16 December 2020
Available online 30 December 2020

(Fan et al., 2014) that could have significant ramifications for the patients' health. In addition, degradation of voice signals produces an acoustic mismatch between the training and operating conditions in automatic PD detection. One possible solution to deal with degraded signals during operation is to use a "multi-condition" training strategy in which the classifier is trained on data with a variety of degradation types at different noise levels. Even though this strategy has proven successful for some speech-based applications such as automatic speech recognition (Ming, 2004) and speaker recognition (Ming et al., 2007), and making them more robust to noisy environments, there are two major issues associated with multi-condition training for PD detection systems: first, there is no guarantee that the classifier learns the differences in the recording environment instead of the differences between PD and healthy voice; and second, the system may behave unpredictably when a new, unseen degradation type is observed in operation. Alternative solution is to reduce the acoustic mismatch between training and operating conditions. A variety of techniques have been developed for compensating this type of mismatch in different speech-based applications (Gong, 1995; Fakhry et al., 2018; Hansen et al., 2014; Alam et al., 2017; Mammone et al., 1996; Nercessian et al., 2016; Poorjam et al., 2016) which can, in general, be categorized into four classes: (1) searching for robust features which parametrize speech regardless of degradations; (2) transforming a degraded signal to the acoustic condition of the training data using a signal enhancement algorithm[2]; (3) compensating the effects of degradation in the feature space by applying feature enhancement; and (4) transforming the parameters of the developed model to match the acoustic conditions of the degraded signal at operating time. To the best of the authors' knowledge, there is a lack of studies of the impact of acoustic mismatch and the effect of compensation on the performance of PD detection systems. Vasquez-Correa et al. proposed a pre-processing scheme by applying a generalized subspace speech enhancement technique to the voiced and unvoiced segments of a speech signal to address the PD detection in non-controlled noise conditions (Vasquez-Correa et al., 2015). They showed that applying speech enhancement to the unvoiced segments leads to an improvement in detection accuracy while the enhancement of voiced segments degrades the performance. However, this study is limited in terms of degradation types as it only considered the additive noise. Moreover, they only evaluated the impact of an unsupervised enhancement method on PD detection performance, while the supervised algorithms have, in general, shown to reconstruct higher quality signals as they incorporate more prior information about the speech and noise.

Another open question which, to the authors' knowledge, has not been addressed is whether applying "appropriate" signal enhancement algorithms to the degraded signals will result in an improvement in PD detection performance. Answering this question, however, requires prior knowledge about the presence and type of degradation in voice signals, which can be achieved by controlling the quality of recordings prior to analysis. Quality control of the voice recordings is typically performed manually by human experts which is a very costly and time consuming task, and is often infeasible in online applications. In Poorjam et al. (2019a), the problem of quality control in remote speech data collection has been approached by identifying the potential outliers which are inconsistent, in terms of the quality and the context, with the majority of speech samples in a data set. Even though very effective in finding outliers, it is not capable of detecting the type of degradation nor identifying short-term protocol violations in recordings. To identify the type of degradation in pathological voices, Poorjam et al. proposed two different parametric and non-parametric approaches to classify degradations commonly encountered in remote pathological voice analysis into four major types, namely background noise, reverberation, clipping and coding (Poorjam et al., 2017, 2018b). However, the performance of these approaches is limited when new degradation types are introduced. Furthermore, the presence of outlier recordings, which do not contain relevant information for PD detection due to long-term protocol violations, is not considered in these methods and, therefore, there is no control over the class assignment for such recordings. To address the frame-level quality control in pathological voices, Badawy et al. proposed a framework for detecting short-term protocol violations using a nonparametric switching autoregressive model (Badawy et al., 2018). In Poorjam et al. (2019b), a highly accurate approach for identifying short-term protocol violations in PD voice recordings has been proposed which fits an infinite hidden Markov model to the frames of the voice signals in the mel-frequency cepstral domain. However, these two approaches do not identify short-term degradations (e.g. the presence of an instantaneous background noise) in voice signals.

To overcome the explained limitations in the existing methods, we propose two approaches for controlling the quality of pathological voices at recording-level and frame-level in this paper. In the recording-level approach, separate statistical models are fitted to the clean voice signals and the signals corrupted by different degradation types. The likelihood of a new observation given each of the models is then used to determine its degree of adherence to each class of acoustic conditions. This gives us the flexibility not only to associate multiple classes to a voice signal corrupted by a combination of different degradations, but also to consider a recording as an outlier or a new degradation when it is rejected by all the models. In the frame-level approach, on the other hand, we extend the work in Poorjam et al. (2019b) to identify short-term protocol violations and degradations in voice signals at the same time. We show how these quality control approaches can effectively inform the choice of signal enhancement methods and, consequently, improve the PD detection performance. The contribution of this paper is thus three-fold: (1) we investigate the impact of acoustic mismatch between training and operating conditions, due to degradation in test signals, on the PD detection performance; (2) to identify this mismatch, we propose two different approaches to automatically control the quality of pathological voices at frame- and recording-level; and (3) to efficiently reduce this mismatch, given that the specific degradation is known, we explore a variety of state-of-the-art enhancement algorithms and their effectiveness in improving the performance of a PD detection system. The rest of the paper is organized as follows. In Section 2, we describe the structure of data we used in different experiments in the paper. Section 3 explains the PD detection system that we have used for the experiments throughout this paper. In Section 4, we investigate the impact of three major types of signal degradation commonly encountered in remote voice analysis, namely noise, reverberation and nonlinear distortion, on the performance of the PD detection system. Following that, in Section 5, we investigate the influence of noise reduction, dereverberation, and declipping algorithms on the performance of the PD detection system. In Section 6, we propose two different quality control approaches and investigate how these methods can improve the performance of PD detection. Finally, Section 7 summarizes the paper.

2. Data structure

Before we start the analyses, it is worth explaining the data structure we used in the series of experiments in this paper. As illustrated in Fig. 1, we use the mPower mobile Parkinson's disease (MMPD) data set (Bot et al., 2016) which consists of more than 65,000 iPhone recordings of the sustained vowel /a/ phonations by PD patients and healthy speakers of both genders from the US. The mean ± standard deviation (STD) of the duration of the data set is 10 ± 0.1 s. The designed voice test protocol for this data set required the participants to hold the phone in a similar position to making a phone call, take a deep breath and utter a sustained vowel /a/ at a comfortable pitch and intensity for 10 s. From this data set, we selected three disjoint subsets: (1) a subset of 800 good-quality voice samples containing 400 PD patients and 400 healthy controls selected equally from both genders, (2) a subset of 8000 random samples equally from both genders and

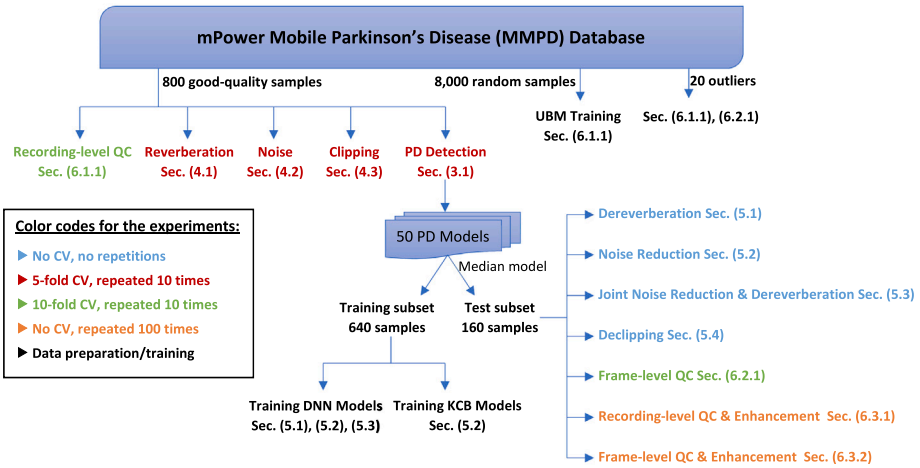


Fig. 1. The data structure used in the experiments. CV and QC stand for cross-validation and quality control, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

target classes, and (3) a subset of 20 outlier samples which contain irrelevant sounds for PD detection such as the sound of a dog barking or a laughter. The names of the experiments and the sections in which each subset is used are indicated in the figure. It should be noted that due to the inherent interference of recording equipment, it is nearly impossible to obtain a perfectly clean and distortion-free voice sample, even if it is captured with a high-quality microphone and in a noise-free, anechoic chamber. Therefore, by the terms "good-quality" or "clean" recording, used interchangeably in this paper, we mean a voice sample in which no ambient noise, reverberation or distortion is perceived and the recording fully complies with the test protocol. The quality of the samples in the first subset is evaluated by manually inspecting the recordings.

In addition to the MMPD data set, we used two noise data sets. The first one contains 4 types of noise, namely babble, restaurant, office, and street noise.[3] These noise types, typically encountered in real situations, are used in the experiment designed for investigating the effect of noise on the PD detection performance, and in the experiments designed to evaluate the performance of quality control algorithms. These experiments can be found in Sections 4.2, 6.1, 6.2, and 6.3. The second noise data set is a subset of NOISEX-92 database (Varga and Steeneken, 1993) and contains 3 types of noise, namely babble, factory, and F16 noise. The NOISEX-92 is commonly used in speech enhancement studies to design and evaluate the noise reduction algorithms. We also use this data set in the experiments designed for the impact of noise reduction on the PD detection performance (Sections 5.2 and 5.3). To make sure that there is no overlap in noise types in the degradation detection experiments, we select a random segment of a noise file and add it to the clean signal.

For experiments that involve reverberation, we use two different room impulse responses (RIRs) to reverberate the clean signals: measured and synthetic RIRs. The measured RIRs, sampled from the Aachen Impulse Response (AIR) database (Jeub et al., 2009), are measured with a dummy head and a mock-up phone in different locations of a wide variety of realistic indoor environments such as an office room, a lecture room, a stairway, and a corridor. The simulated RIRs, on the other hand, are artificially generated by the RIR Generator toolbox (Habets, 2006) based on the parameters of the acoustic environment and the position of a speaker and a microphone. While the measured RIRs provide more realistic reverberations, the RIR generator gives more flexibility to control over the reverberation time of the RIRs. For this reason, we used the synthetic RIRs to investigate the impact of reverberation and dereverberation on the PD detection performance in Section 5, and used the measured RIRs for degradation detection experiments in Section 6.

Due to the randomness involved in some experiments, we repeated the experiments to obtain the distribution of the metrics. The number of iterations, indicated by different colors in Fig. 1, depends on the computational complexity of the algorithms used in each experiments. Moreover, for the experiments that investigate the acoustic mismatch between training and test conditions, we use a PD detection model which is trained on the clean data. In these experiments, indicated by blue and orange colors in the figure and can be found in Sections 5 and 6, we do not use cross-validation for evaluation. For other experiments (in Sections 3, 4, 6.1, and 6.2), we applied the cross-validation, and the number of folds depend on the computational complexity of algorithms.

3. Parkinson's disease detection system

The problem of PD diagnosis from voice has been addressed by many researchers which, in a broad sense, can be categorized into two categories: (1) the regression-based approaches (Tsanas et al., 2010, 2011), which map the dysphonia measures to a clinical score measuring PD symptom severity, such as the unified Parkinson's disease rating scale (UPDRS) (Ramaker et al., 2002), using regression analysis methods; and (2) the classification-based methods (Tsanas et al., 2012; Moro-Velázquez et al., 2018; Orozco-Arroyave et al., 2016), which distinguish between PD patients and healthy speakers. The former approaches provide more clinically useful information and facilitate the monitoring of the PD symptoms progression in individuals. The development of the classification-based approaches, on the other hand, do not necessarily require the severity scores, which are not always available and are shown to be very noisy to accurately regress on them (Evers et al., 2019). This makes the classification-based approaches useful for a quick screening of the population to provide a short list of the PD patients for further inspections. Since the main focus of this paper is to study the influence of the quality control and enhancement on the performance of PD detection systems, we do not propose a new PD detection algorithm. Instead, in this study, we concentrate only on the classification-based approaches, and choose one of the recently proposed algorithms and use it for further quality control and enhancement experiments. However, as an important class of PD diagnosis algorithms, the future work could focus on investigating the effect of the quality control algorithms on the regression-based approaches.

[3] The babble, restaurant and street noise files have been taken from https://www.soundjay.com/index.html and the office noise has been taken from https://freesound.org/people/DavidFrbr/sounds/327497.

The PD detection approach we use in this study was proposed by Moro-Velázquez et al. in Moro-Velázquez et al. (2018). In this method, Gaussian mixture models (GMMs) are fitted to the frames of the voice recordings of the PD patients and the healthy controls (HC) parametrized by perceptual linear predictive (PLP) coefficients (Hermansky, 1990). The authors in Moro-Velázquez et al. (2018) used PLP parametrization since perceptual features have been shown to have more discriminative power in PD detection than conventional, clinically interpretable, features (such as standard deviation of fundamental frequency, jitter, shimmer, harmonic-to-noise ratio, glottal-to-noise excitation ratio, articulation rate, and frequencies of formants), particularly when the voice is more noise-like, aperiodic, irregular and/or chaotic, which typically occurs in more advanced stages of PD (Orozco-Arroyave et al., 2013; Brabenec et al., 2017; Mekyska et al., 2016). Moro-Velázquez et al. showed in Table 5 of Moro-Velázquez et al. (2018) that PLP parametrization can, on average, achieve a better performance than mel-frequency cepstral coefficient (MFCC) for different speech materials. Moreover, the perceptual analysis of different vowels in the study by Orozco-Arroyave et al. (2013) suggests that the PLP coefficients can better parametrize the vowel /a/ than other perceptual features. Given that the speech material for the PD detection experiments in this paper is the sustained vowel /a/, we also use PLP parametrization in our experiments.

Acoustic features of the PD patients' recordings and those of the healthy controls are modeled by GMMs with the likelihood function defined as:

$$p(\boldsymbol{x}_t|\lambda) = \sum_{c=1}^{C} b_c p(\boldsymbol{x}_t|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{1}$$

where $\boldsymbol{x}_t$ is the feature vector at time frame t, b_c is the mixture weight of the cth mixture component, C is the number of Gaussian mixtures, $p(\boldsymbol{x}_t|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ is a Gaussian probability density function where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the mean and covariance of the cth mixture component, respectively. The parameters of the model, $\lambda = \{b_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{C}$, are trained through the expectation–maximization algorithm (Reynolds and Rose, 1995).

Given $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$, a sequence of feature vectors, the goal in PD detection is to find the model which maximizes $p(\lambda_j|\boldsymbol{X})$, where $j \in \{\text{PD}, \text{HC}\}$. Using the Bayes' rule, independence assumption between frames, and assuming equal priors for the classes, the PD detection system computes the log-likelihood ratio for an observation as:

$$\sigma(\boldsymbol{X}) = \sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_{\text{PD}}) - \sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_{\text{HC}}). \tag{2}$$

The final decision about the class assignment for an observation is made by setting a threshold over the obtained score.

3.1. Experimental setup

In this study, we use the sustained vowel /a/ as the speech material for PD detection since they provide a simpler acoustic structure to characterize the glottal source and resonant structure of the vocal tract than running speech. We consider the mPower mobile Parkinson's disease (MMPD) data set (Bot et al., 2016), described in Section 2. To evaluate the performance of the PD detection system under matched acoustic conditions, a subset of 800 good-quality voice samples, consisting of 400 PD patients and 400 healthy controls equally from both genders, have been selected from this data set. It is worth mentioning that since the health status in this data set is self-reported, to have more reliable samples for the PD class, we selected participants who self-reported to have PD, claimed that they have been diagnosed by a medical professional with PD, and recorded their voice right before taking PD medications. For the healthy control class, we selected participants who self-reported being healthy, do not take PD medications, and claimed that they have not been diagnosed by a medical professional with PD. All speakers of this subset had an age range of 58 to 72. The mean ± STD of the age of PD patients and healthy controls are 64 ± 4 and 66 ± 4, respectively. For all experiments in this paper, we downsampled the recordings from 44.1 kHz to 8 kHz since the enhancement algorithms used in this work are operating at 8 kHz. To extract the PLP features, voice signals are first segmented into frames of 30 ms with 10 ms overlap using a Hamming window. Then, 13 PLP coefficients are computed for each frame of a signal. To consider the dynamic changes between frames due to the deviations in articulation, a first- and a second-order orthogonal polynomials are fitted to the two feature vectors to the left and right of the current frame. These features, which are referred to as *delta* and *double-delta*, were appended to the feature vector to form a 39-dimensional vector per each frame. The number of mixture components for the GMMs was set to 32.

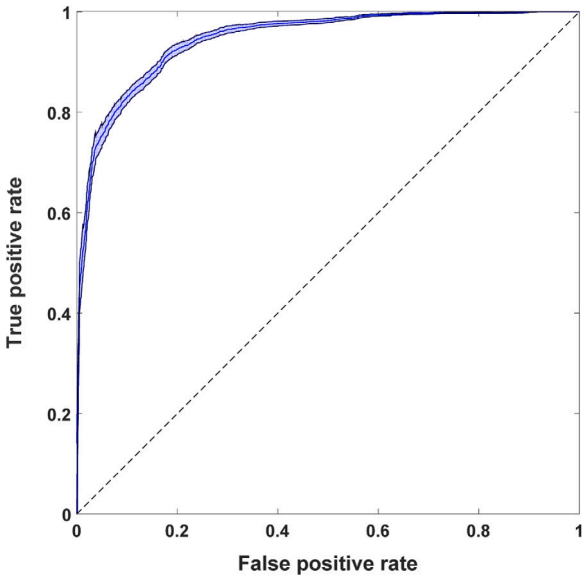
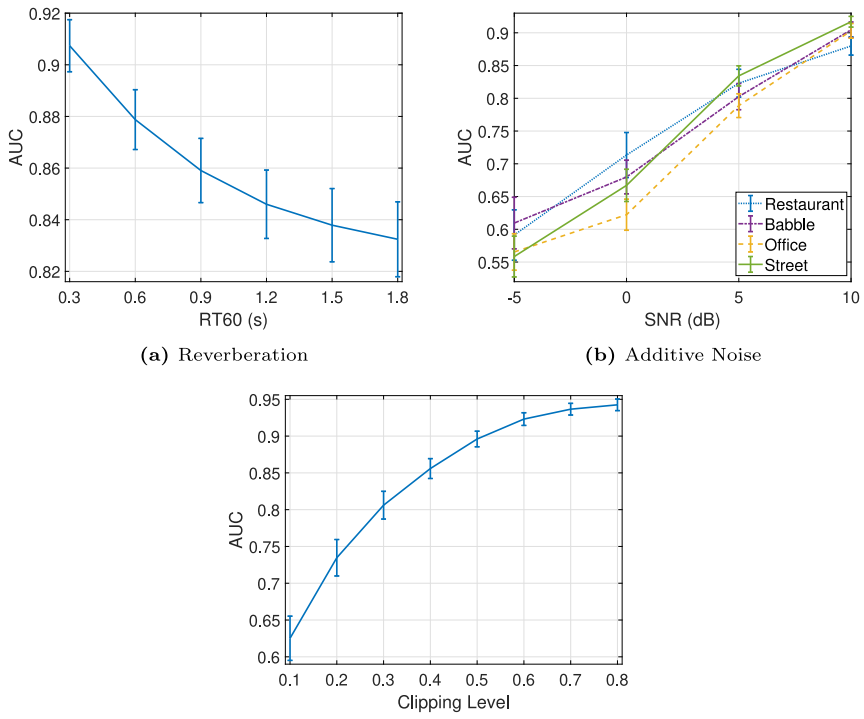


Fig. 2. The ROC curve of the PD detection system, along with 95% confidence interval shaded in blue. The dashed line shows the chance level.

3.2. Results

To evaluate the performance of the PD detection system in a matched acoustic condition, we used 5-fold cross validation (CV) in which the recordings were randomly divided into 5 non-overlapping and equal sized subsets. Since we only used one recording per speaker, there is no risk of finding recordings of the same speaker in both training and test subsets. The entire CV procedure was repeated 10 times to obtain the distribution of detection performance. Fig. 2 shows the performance in terms of the receiver operating characteristic (ROC) curve, along with 95% confidence interval. The ROC is a probability curve which plots the true positive rate against the false positive rate for different decision thresholds. The area under the curve (AUC) summarizes the ROC curve and represents the performance of a detection system by a single number between 0 and 1; the higher the performance, the closer the AUC value is to 1. Comparing with the commonly used classification accuracy, defined as the percentage of correct predictions, the AUC is the preferred metric in this paper since it indicates how well the model can distinguish between two classes which sets a fundamental limit to the classification accuracy metric. Moreover, the AUC is independent of the decision threshold, which is a user- and application-specific parameter, whereas the estimation of the accuracy requires a threshold over the scores. The mean AUC for this PD detection system is 0.95.



(a) Reverberation (b) Additive Noise

(c) Nonlinear Distortion

Fig. 3. Performance of the PD detection system in acoustic mismatch conditions due to different degradations in test signals in terms of AUC, along with 95% confidence intervals.

4. Impact of signal degradation on PD detection

The PD detection system explained in the previous section gave a mean AUC of 0.95 in a matched acoustic condition. That is, when it was trained and tested using the clean recordings. However, as alluded to in the introduction, recordings collected remotely in an unsupervised manner are seldom clean as they are often degraded by different types of degradation. In this section we investigate the effect of acoustic mismatch between training and operating conditions on the performance of the PD detection system. To this aim, we artificially degrade the test signals using three types of degradation commonly encountered in remote voice analysis, namely reverberation, background noise and nonlinear distortion. It should be noted that, even though we tried to choose the most reliable samples from the MMPD data set, the labels are not 100% reliable as the diagnosis is self-reported. For this reason, we are more interested in how the relative PD detection performance is influenced systematically under the application of different experimental conditions.

4.1. Reverberation

Reverberation is a phenomenon that occurs when the signal of interest is captured in an acoustically enclosed space. Apart from the direct component, the microphone receives multiple delayed and attenuated versions of the signal, which is characterized by the room impulse response (RIR). A metric commonly used to measure the reverberation is the reverberation time (RT60) (Vorländer, 2007). The presence of reverberation has shown to degrade the performance of speech-based applications such as speech and speaker recognition (Yoshioka et al., 2012; Castellano et al., 1996). In this section, we investigate the effect of reverberation on the PD detection performance. To this aim, we used 5-fold CV repeated 10 times to evaluate the performance. In each iteration, the model was trained using the clean recordings of the training subset, and evaluated on the recordings of the disjoint test subset which were filtered with synthetic RIRs of RT60 varying from 300 ms to 1.8 s in 300 ms steps measured at a fixed position in a room of dimension 10 m × 6 m × 4 m. The distance between source and microphone is set to 2 m. The room impulse responses were generated using the image method (Allen and Berkley, 1979) and implemented using the RIR Generator toolbox (Habets, 2006). Fig. 3a shows the impact of reverberation on the PD detection performance in terms of the mean AUC along with 95% confidence intervals. We can observe from the plot that the PD detection system exhibits lower performance in reverberant environments, as expected, and the amount of degradation is related to the RT60.

4.2. Background noise

Background noise is one of the most common types of degradation occurring during remote voice analysis. In this section we restrict ourselves to additive background noise and investigate how this can influence the PD detection performance. To this aim, we performed the same CV procedure used for evaluating the impact of reverberation (explained in the previous section). In each iteration, the model was trained using the clean recordings of the training subset, and evaluated using the recordings of the test subset contaminated by an additive noise. The entire procedure was repeated for four different noise types, namely babble, restaurant, office and street noise, selected from the first noise data set (explained in Section 2). To choose a more realistic range of the signal-to-noise ratio (SNR) values for this experiment, we applied the waveform amplitude distribution analysis (WADA) algorithm (Kim and Stern, 2008) to the entire signals of the MMPD data set to roughly estimate the global SNR of the signals in a remotely collected data set. Even though we discussed in Poorjam et al. (2018a) that the SNR estimation algorithms, such as WADA, that are developed for normal speech, are not highly accurate in estimating the SNR in pathological voices, it gives a rough idea of the range of SNR values in this data set. Considering the distribution of SNRs, illustrated in Fig. 4, and accounting for an error in the SNR estimation by WADA, the range [−5,10] dB can be considered realistic in a remotely collected data set. Therefore we contaminated the signals under different SNR conditions ranging from −5 dB to 10 dB in 5 dB steps. Fig. 3b illustrates the

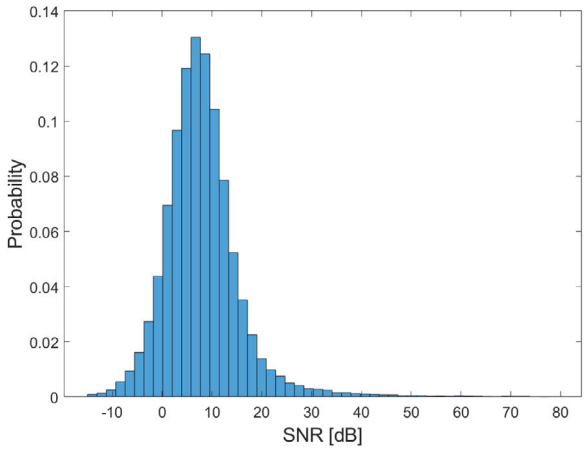


Fig. 4. The normalized histograms of the estimated global SNR values for the entire MMPD databases using the waveform amplitude distribution analysis algorithm (Kim and Stern, 2008).

impact of different noise types and different SNR conditions on the performance of the PD detection system in terms of the mean AUC along with the 95% confidence intervals. We can observe a similar trends for all noise types that the PD detection performance decreases as the noise level increases.

4.3. Clipping

In remote voice analysis, nonlinear distortion can manifest itself in speech signals in many different ways such as clipping, compression, packet loss and combinations thereof. Here, we consider hard clipping, or magnitude saturation, as an example of nonlinear distortion in signals which is caused when a signal fed as an input to a recording device exceeds the dynamic range of the device (Eaton and Naylor, 2013). By defining the clipping level as a proportion of the unclipped peak absolute signal amplitude to which samples greater than this threshold are limited, we can investigate the impact of clipping on the PD detection performance. To this aim, the clean recordings of the test subset in each iteration of the CV were clipped with different clipping levels ranging from 0.1 to 0.8 in 0.1 steps. Fig. 3c shows the performance as a function of clipping level. Similar to the other types of degradation, it can be observed that increasing the distortion level in voice signals decreases the PD detection performance.

5. Impact of signal enhancement on PD detection performance

As seen in Section 4, the degradation introduced to the signals can lead to reduction in the performance of the PD detection system. Since there are practically an infinite number of possible types and combinations of nonlinear distortion that can be present in a signal, and since there is a lack of well-documented algorithms for dealing with most of the distortions (even in isolation), in this section, we only consider the degradations for which there are well-documented and verified enhancement algorithms such as noise reduction, dereverberation, and declipping and investigate the effects of these algorithms on the PD detection performance. To this end, from the 50 PD detection models developed and evaluated through 10 iterations of the 5-fold cross-validation procedure, as explained in Section 3.2, we selected one of the two models which showed the median performance and used it for further enhancement experiments in this section. We have used a total of 160 recordings for testing the algorithms used in this section. We will restrict ourselves to single channel enhancement algorithms. There exist a variety of objective and subjective metrics to measure the quality of the enhanced speech signal such as SNR, signal-to-distortion ratio (Vincent et al., 2007), perceptual evaluation of speech quality (Hu and Loizou, 2007) and short-time objective intelligibility (Taal et al., 2010). However, since our main goal in this work is to study the influence of speech enhancement on the PD detection performance, we evaluate the effectiveness of the algorithms in terms of the AUC.

5.1. Dereverberation

Some of the popular classes of dereverberation techniques are the spectral enhancement methods (Habets, 2007), probabilistic model based methods (Jukić and Doclo, 2014; Jukić et al., 2015; Jeub et al., 2010), and inverse filtering based methods (Kameoka et al., 2009; Huang and Benesty, 2003). Spectral enhancement methods estimate the clean speech spectrogram by frequency domain filtering using the estimated late reverberation statistics. The probabilistic model based algorithms model the reverberation using an autoregressive (AR) process, and the clean speech spectral coefficients using a certain probability distribution function. The estimated parameters of the model are then used to perform dereverberation. Lastly, the inverse filtering methods use a blindly estimated room impulse response to design an equalization system. These methods, which are mainly developed for running speech, assume that the signal at a particular time–frequency bin is uncorrelated with the signals at that same frequency bin for frames beyond a certain number (Jukić et al., 2015). However, this assumption is not valid for the sustained vowels which makes the dereverberation of the sustained vowels more challenging. Recently, deep neural network (DNN) based dereverberation algorithms have gained attention (Han et al., 2015; Santos and Falk, 2018) since they relax the assumption of uncorrelated neighboring time–frequency bins. The underlying principle of the DNN-based methods is to train a DNN to map the log-magnitude spectrum of the degraded speech to that of the desired speech.

In this section, we investigate the effectiveness of different dereverberation algorithms in improving the PD detection performance. For dereverberation experiments, we used three different algorithms: a probabilistic model based algorithm proposed in Jukić et al. (2015) (denoted as WPE-CGG, weighted prediction error with complex generalized Gaussian prior), an algorithm based on the inverse filtering of the modulation transfer function (Kameoka et al., 2009) (denoted as IF-MU, inverse filtering with multiplicative update), and a DNN-based algorithm proposed in Han et al. (2015) (denoted as DNNSE-R, deep neural network speech enhancement for reverberant signals). It should be noted that the WPE-CGG and the IF-MU are unsupervised methods whereas the DNNSE-R is a supervised method. For the DNN-based algorithm, a feedforward neural network with 3 hidden layers of 1600 neurons was used. To take into account the temporal dynamics, features of 11 consecutive frames (including the current frame, 5 frames to the left and 5 frames to the right over time) were provided to represent the input features of the current frames. The parameters of the neural network are optimized by minimizing the mean square error loss function. For more detail about the network architecture and phase estimation for signal reconstruction, see Han et al. (2015). To train the DNN model, we selected 640 clean recordings from the MMPD data set and filtered them with the synthetic room impulse responses of RT60 ranging from 200 ms to 1 s in steps of 100 ms using the implementation in Habets (2006) for a particular source and receiver position in a room of dimensions 10 m × 6 m × 4 m. For testing, the position of the receiver was fixed while the position of the source was varied randomly from 60 degrees left of the receiver to 60 degrees right of the receiver. Fig. 5 shows the performance of the PD detection in terms of AUC for the different dereverberation algorithms. It can be observed from the figure that only DNNSE-R is able to improve the PD detection performance while the other two methods degrade the performance. This is mainly due to two reasons: first, the DNNSE-R is a supervised algorithm while the WPE-CGG and IF-MU are unsupervised; and second, the underlying assumption of the two unsupervised algorithms does not hold for the sustained vowels. We have also included the case of zero RT60 to investigate the impact of processing of the clean recordings by these dereverberation algorithms.

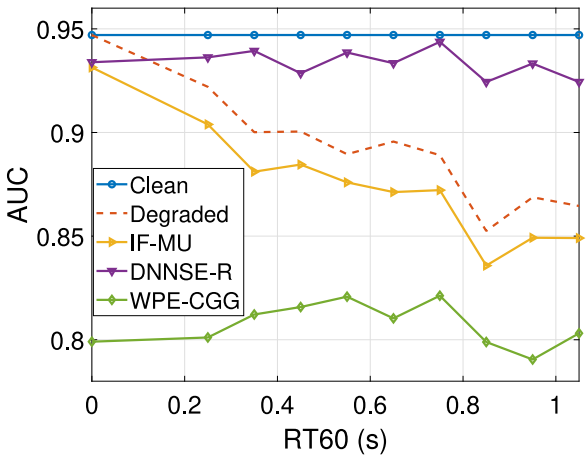


Fig. 5. Impact of different dereverberation algorithms on the PD detection performance, in terms of AUC.

5.2. Noise reduction

Methods for performing noise reduction can be broadly categorized into supervised and unsupervised methods. Unsupervised methods do not assume any prior knowledge about identity of the speaker or noise environment. The supervised methods, on the other hand, make use of training data to train the models representing the signals of interest or the noise environment. Some of the popular classes of supervised speech enhancement methods include the codebook-based methods (Srinivasan et al., 2006; He et al., 2017), non-negative matrix factorization based methods (Mohammadiha et al., 2013; Fakhry et al., 2018) and the DNN-based methods (Wang and Chen, 2018). In the supervised algorithms, the speech and noise statistics/parameters estimated using the training data are exploited within a filter to remove the noise components from the noisy observation. In this section, we used two supervised methods and one unsupervised method to investigate the effect of different noise reduction algorithms in reducing the acoustic mismatch between training and operating conditions.

The first supervised algorithm is based on the framework proposed in Kavalekalam et al. (2019). In this approach, a Kalman filter, which takes into account the voiced and unvoiced parts of speech (Goh et al., 1999), is used for enhancement. The filter parameters consist of the AR coefficients and excitation variance corresponding to speech and noise along with the pitch parameters (i.e. the fundamental frequency and the degree of voicing). Based on Kavalekalam et al. (2019), the AR coefficients and excitation variance of the speech and noise are estimated using a codebook-based approach, and the pitch parameters are estimated from the noisy signal using a harmonic model based approach (Nielsen et al., 2017). We refer to this method in the rest of this paper as the Kalman-CB. This algorithm has been selected because of its good performance in noise reduction in terms of quality and intelligibility based on both objective and subjective measures. The speech codebook was trained using 640 clean recordings selected from the MMPD data set (equally from both genders). To train the noise codebook, we used babble, restaurant, office and street noises to create four sub-codebooks. During the testing phase, all sub-codebooks, except the one corresponding to the target noise, were concatenated to form the final noise codebook. The size of the speech and noise codebooks were set to 8 and 12, respectively.

The second supervised enhancement method is the DNN-based algorithm proposed in Han et al. (2015). This algorithm is the same as the one we used for dereverberation experiments, except it is trained using the noisy signals. We refer to this method in the rest of this paper as the DNNSE-N, deep neural network speech enhancement for noisy signals. This algorithm has been selected because, besides improvements in objective measures, it showed improvement in performance of automatic speech recognition in noisy environments. To train the DNNSE-N, we used the same 640 clean recording that we used for training the speech codebook in the Kalman-CB algorithm. The recordings were contaminated by three types of noise, namely babble, factory and F16 noises taken from NOISEX-92 database (Varga and Steeneken, 1993) under different SNR conditions selected randomly from the continuous interval [0,10] dB.

We used, as an unsupervised speech enhancement method, the algorithm proposed in Erkelens et al. (2007) which is based on the minimum mean-square error (MMSE) estimation of discrete Fourier transform (DFT) coefficients of speech while assuming a generalized gamma prior for the speech DFT coefficients. This method, denoted as MMSE-GGP, is a popular unsupervised algorithm which uses the MMSE-based tracker for noise power spectral density estimation.

Fig. 6 shows the impact of the noise reduction algorithms on the PD detection performance in terms of AUC for different noise types and SNR conditions. It can be observed from the figures that enhancing the degraded voice signals with the supervised methods in general improves the performance. For instance, applying the Kalman-CB algorithm resulted in 21.3%, 18.3%, 11.1%, and 2.2% relative improvements in the AUC (averaged over 4 different noise types) for the −5 dB, 0 dB, 5 dB, and 10 dB scenarios. However, the unsupervised method shows improvement only in the low SNR range and degrades the PD detection performance in higher SNR scenarios. The low performance of the unsupervised algorithm can be due to the fact that noise statistics in this case is estimated using a method proposed in Gerkmann and Hendriks (2012) which has been designed for running speech rather than the sustained vowels. This observation is somewhat consistent with the statement in Vasquez-Correa et al. (2015), which suggested that applying an unsupervised enhancement algorithm to the voiced segments results in a degradation in PD detection performance.

5.3. Joint noise reduction and dereverberation

In Sections 5.1 and 5.2, we showed the impact of noise reduction and dereverberation when one of these degradations was present in the signal. However, in some cases, the recordings may be degraded simultaneously by reverberation and background noise. There have been methods proposed for joint noise reduction and dereverberation with access to multiple channels (Habets et al., 2008; Kodrasi and Doclo, 2016).

Since we have restricted ourselves to single channel enhancement methods, and motivated by the improvement in the PD detection performance as a result of using the DNN-based algorithm for noise reduction and dereverberation, in this section, we investigate the effectiveness of this algorithm in performing joint noise reduction and dereverberation. In this case, the input to the DNN is the log-magnitude spectrum of the signal which is degraded by reverberation and background noise. This method is referred to, in the rest of the paper, as DNNSE-NR, deep neural network speech enhancement for noisy and reverberant signals. For training the DNN model, the same 640 clean recordings that we used in the previous enhancement experiments were filtered with RIRs of different RT60s ranging from 400 ms to 1 s with 200 ms steps. Then, three types of noise, namely babble, factory and F16 noises (taken from NOISEX-92 database) were randomly added to the reverberant signals at different SNRs selected uniformly at random from the continuous interval [0,10] dB. Table 1 summarizes the impact of joint noise reduction and dereverberation using the DNNSE-NR algorithm on the PD detection performance. In this table, we have also included the cases of infinite SNR and zero RT60 to investigate the effect of the enhancement system when the clean recordings or the ones degraded by only noise or reverberation were processed by this algorithm. It can be observed for the case of babble noise that the DNNSE-NR improves the PD detection performance in most of the cases when reverberation and background noise coexist and in the cases

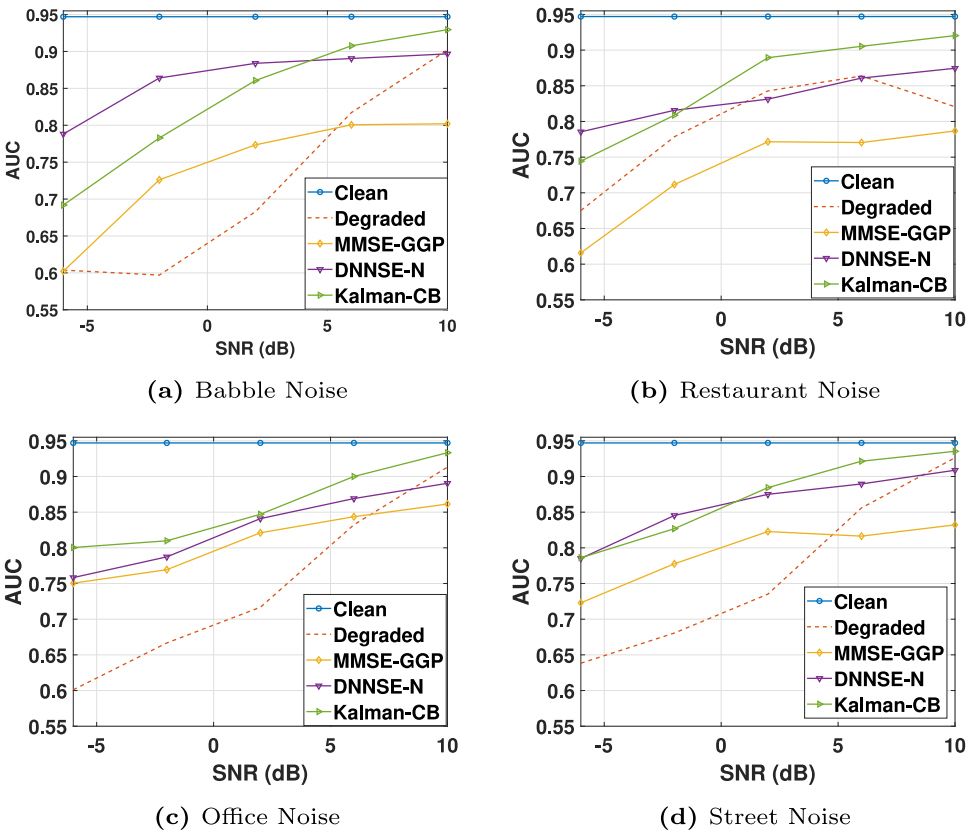


Fig. 6. Impact of different noise reduction algorithms on the PD detection performance, in terms of AUC, under different noise types and SNR conditions.

where only noise is present. However, in the case of only reverberation, the DNNSE-NR shows improvement only in the cases where RT60 is 400 ms and above. It should be noted that the babble noise used for training and testing were taken from two different noise databases. In the case of restaurant noise, improvement in PD detection performance is observed only in the low SNRs, namely −2 dB and −6 dB. The results of the restaurant noise is interesting in a sense that it shows how the DNNSE-NR algorithm can generalize for a noise type not seen during the training phase.

5.4. Declipping

Declipping is the process of restoration of the clipped audio signal by estimating the original signal. Given that the samples of an observed clipped signal are considered as either reliable (unclipped) or clipped samples, the declipping methods can be broadly categorized into two classes: consistent and inconsistent methods. In consistent methods, either or both the reliable samples in the reconstructed signal are equal to those in the clipped signal (referred to as the *reliable part consistency*) or the estimated missing samples in the reconstructed signal hold the clipping constraints defined in the clipping model (referred to as the *clipped part consistency*). In inconsistent methods, on the other hand, the samples of the restored signal in the clipped parts do not require to hold the clipping constraint nor the reconstructed samples in the reliable parts need to be equal the observed clipped signal (Záviška et al., 2020).

To investigate the impact of different declipping algorithms on the performance of the PD detection, we selected one fully consistent method, one fully inconsistent method, and one method that is consistent in the reliable part. The fully consistent approach, proposed by Kitić et al. in Kitić et al. (2015), is a sparsity-based approach which uses the sparse analysis data model and the alternating direction method of multipliers (ADMM) to approximate the optimal solution of the ill-conditioned inverse problem of declipping. We refer to this algorithm in the rest of this paper as the ASPADE, *Analysis* version of *SParse Audio DEclipper*. The algorithm operates sequentially on individual frames of the signal. For more details about the algorithm, we refer to Kitić et al. (2015) and Záviška et al. (2018). The next declipping approach we used is proposed by Siedenburg et al. in Siedenburg et al. (2014). This method uses the iterative thresholding algorithm and the concept of the social sparsity (Kowalski et al., 2013) to consider the temporal dependencies between adjacent Gabor time–frequency coefficients and to approximate a solution to the problem of declipping audio signals. The formulation of this algorithm allows inconsistency of both clipped and reliable parts. This algorithm is referred to as the Social Sparsity in the rest of this paper. For more details about this method see Siedenburg et al. (2014). The last declipping algorithm we used is the method proposed by Janssen et al. in Janssen et al. (1986). This method which is based on AR modeling of audio signals, considers the declipping as a problem of recovering the missing samples of a signal by generating them from the AR model. The Janssen's method is only consistent in the reliable part since it is used to generate the samples in the clipped part by a linear estimation of the unclipped samples in the reliable part.

Fig. 7 shows the impact of different declipping algorithms on the performance of PD detection in terms of the AUC for different clipping levels ranging from 0.2 to 0.8 in 0.2 steps. For these experiments, we used the MATLAB implementation of the algorithms provided by Záviška et al. (2020) and set the declipping models' parameters accordingly. We can observe that all three declipping algorithms used in this experiment could improve the performance of the PD detection when the signals are undergone mild or moderate clipping. However, in case of severe clipping, the ASPADE method outperformed the others and made the PD detection robust against hard clipping.

6. Automatic quality control in pathological voice recordings

We have shown in the previous section that, assuming the specific degradation is known, there exist algorithms to effectively transform

Table 1
Impact of joint noise reduction and dereverberation using the DNN-SE algorithm on the PD detection performance. Bold numbers indicate the improvement in performance.

			Babble noise: SNR (dB)					
			−6	−2	2	6	10	inf
RT60 (s)	0	Degraded	0.67	0.59	0.69	0.80	0.90	0.95
		DNN-SE	**0.80**	**0.89**	**0.89**	**0.89**	0.89	0.91
	0.2	Degraded	0.56	0.64	0.72	0.81	0.89	0.95
		DNN-SE	**0.82**	**0.89**	**0.87**	**0.89**	0.89	0.91
	0.4	Degraded	0.54	0.66	0.70	0.80	0.84	0.90
		DNN-SE	**0.78**	**0.84**	**0.85**	**0.89**	**0.86**	**0.91**
	0.6	Degraded	0.64	0.70	0.71	0.78	0.81	0.88
		DNN-SE	**0.75**	**0.83**	**0.85**	**0.85**	**0.88**	**0.89**
	0.8	Degraded	0.67	0.70	0.73	0.79	0.83	0.89
		DNN-SE	**0.81**	**0.83**	**0.86**	**0.87**	**0.88**	**0.91**
	1	Degraded	0.54	0.68	0.74	0.81	0.84	0.88
		DNN-SE	**0.80**	**0.81**	**0.86**	**0.86**	**0.88**	**0.90**
			Restaurant noise: SNR (dB)					
			−6	−2	2	6	10	inf
RT60 (s)	0	Degraded	0.71	0.81	0.82	0.82	0.90	0.95
		DNN-SE	**0.77**	0.81	0.82	**0.83**	0.87	0.91
	0.2	Degraded	0.67	0.75	0.76	0.85	0.89	0.95
		DNN-SE	**0.74**	**0.79**	**0.79**	0.84	0.87	0.91
	0.4	Degraded	0.62	0.73	0.83	0.83	0.83	0.92
		DNN-SE	**0.73**	**0.77**	0.79	0.83	0.82	0.91
	0.6	Degraded	0.59	0.79	0.81	0.80	0.86	0.89
		DNN-SE	**0.69**	**0.81**	0.79	**0.81**	0.84	**0.91**
	0.8	Degraded	0.58	0.76	0.82	0.81	0.86	0.87
		DNN-SE	**0.75**	0.76	0.80	**0.84**	**0.87**	**0.90**
	1	Degraded	0.65	0.75	0.76	0.82	0.83	0.85
		DNN-SE	**0.76**	0.75	**0.78**	0.81	0.82	**0.90**

a voice signal from a degraded condition into the acoustic condition in which models are trained. Choosing the appropriate enhancement algorithm, however, requires prior knowledge about the presence and type of degradation in a voice signal. In this section, we introduce two approaches to automatically control the quality of recordings. The first approach detects, at recording level, the presence and type of degradation which has influenced the majority of frames of the signal. The second approach, on the other hand, detects short-term degradations and protocol violations in a signal.

6.1. Recording-level quality control

The major limitation of the multi-class classification-based approaches for identifying the type of degradation in a voice signal (Poorjam et al., 2017, 2018b) is that they do not consider the fact that a recording can be subject to an infinite number of possible combinations of degradations in real scenarios. This causes some problems when a signal is contaminated by a new type of degradation for which the classifier has not been trained. Moreover, there is no control in class assignment for a high-quality outlier which do not comply with the context of the data set.

To overcome these limitations, instead of using a multi-class classifier, we propose to use a set of parallel likelihood ratio detectors for the major types of degradations commonly encountered in remote voice analysis, each detecting a certain degradation type. This way, the likelihood ratio statistics of an observation given each of the models can be translated to the degree of contribution of each degradation to the degraded observation. Moreover, completely new degradation types and outliers can be detected if all models reject those observations according to a pre-defined threshold.

In this approach, the task of each detector is to determine whether a feature vector of the time frame t of a voice signal, $\boldsymbol{x}_t$, was contaminated by the corresponding degradation, H_0, or not, H_1. The decision

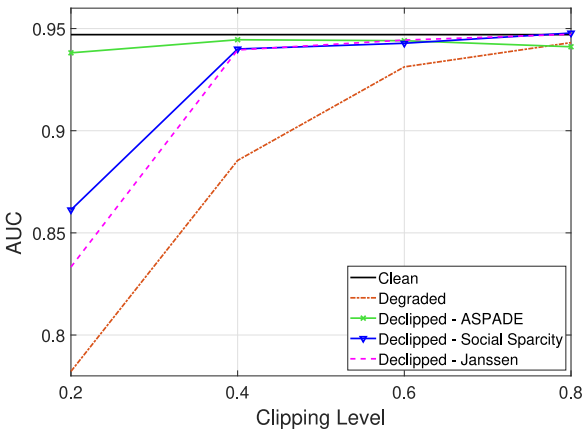


Fig. 7. Impact of various declipping algorithms on the PD detection performance, in terms of AUC under different clipping levels.

about the adherence of each frame of a given speech signal to the hypothesized degradation is then computed as:

$$\log p(\boldsymbol{x}_t|H_0) - \log p(\boldsymbol{x}_t|H_1) \begin{cases} \geq \omega, & \text{accept } H_0 \\ < \omega, & \text{reject } H_0, \end{cases} \tag{3}$$

where ω is a pre-defined threshold for detection, and $p(\boldsymbol{x}_t|H_0)$ and $p(\boldsymbol{x}_t|H_1)$ are respectively the likelihood of the hypotheses H_0 and H_1 given $\boldsymbol{x}_t$.

To model the characteristics of each hypothesized degradation, we propose to fit a GMM of the likelihood function defined in (1) to the frames of the recordings in the feature space. The motivation for using GMMs is that they are computationally efficient models that are capable of modeling sufficiently complex densities as a linear combination of simple Gaussians. Thus, the underlying acoustic classes of the signals might be modeled by individual Gaussian components. While the hypothesized degradation models can be well characterized by using training voice signals contaminated by the corresponding degradation, it is very challenging to model the alternative hypothesis as it should represent the entire space of all possible negative examples expected during recognition. To model the alternative hypothesis, instead of using individual degradation-specific alternative models, we train a single degradation-independent GMM using a large number of clean, degraded and outlier voice signals. Since this background model is used as an alternative hypothesis model for all hypothesized degradations, it is referred to as a universal background model (UBM).

When the UBM is trained, a set of degradation-dependent GMMs are derived for modeling clean, noisy, reverberant and distorted recordings, $D = \{\lambda_d\}_{d=1}^{4}$, by adapting the parameters of the UBM through a *maximum a posteriori* estimation and using the corresponding training data. Given the UBM, λ_{ubm}, and the dth trained degradation model, λ_d, and assuming that the feature vectors are independent, the log-likelihood ratio for a test observation, $\boldsymbol{X}_{\text{ts}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$, is calculated as:

$$\sigma_d(\boldsymbol{X}_{\text{ts}}) = \frac{1}{T}\left(\sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_d) - \sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_{\text{ubm}})\right). \tag{4}$$

The scaling factor in (4) is used to make the log-likelihood ratio independent of the signal duration and to compensate for the strong independence assumption for the feature vectors (Reynolds et al., 2000). The decision for the test observation can be made by setting a threshold over the scores.

To parametrize the recordings for the purpose of degradation detection, we propose to use mel-frequency cepstral coefficients (MFCCs) (Deller et al., 2000). The main motivation for choosing a different speech parametrization for degradation detection than that used for

PD detection is that not only do the MFCCs convey information about the speech context, but also they encode the presence and the level of degradation in signals due to their sensitivity to small changes in signal characteristics caused by degradation (Poorjam et al., 2017; Narayana and Kopparapu, 2009; Pan and Waibel, 2000; Poorjam et al., 2018a). We have demonstrated in Poorjam et al. (2017, 2018a) that degradation in speech signals predictably modifies the distribution of MFCCs by changing the covariance of the features and shifting the mean to different regions in feature space, and the amount of change is related to the degradation level.

6.1.1. Experimental setup

For training the UBM, we randomly selected 8,000 recordings from the MMPD data set. To avoid the UBM model to be biased towards the dominant subpopulations, we make the training data balanced over all subpopulations by randomly dividing this subset into 5 equal partitions of 1600 samples. The recordings of the first partition were randomly contaminated by six different types of noise namely babble, street, restaurant, office, white Gaussian and wind noises under different SNR conditions ranging from −10 dB to 20 dB in 2 dB steps. To produce reverberant signals, the recordings of the second partition were filtered by 46 real room impulse responses of the AIR database (Jeub et al., 2009), measured with a mock-up phone in different realistic indoor environments. As an example of non-linearities in signals, the recordings of the third partition were processed randomly by either clipping, coding or clipping followed by coding. The clipping level was set to 0.3, 0.5 and 0.7. We used 9.6 kbps and 16 kbps code-excited linear prediction (CELP) codecs (Schroeder and Atal, 1985). To consider the combination of degradations in signals, the recordings of the fourth partition were randomly filtered by 46 different real RIRs and added to the noises typically present in indoor environments, namely babble, restaurant and office noise at 0 dB, 5 dB and 10 dB. The recordings of the last partition were used without any processing. The last subset also contains some outliers which do not contain relevant information for PD detection.

For adaptation of the degradation-dependent models, a subset of 800 good-quality recordings of PD patients and healthy speakers of both genders were equally selected from the MMPD data set. From this subset, 200 recordings were corrupted by babble, restaurant, street and office noises under different SNR conditions ranging from −5 dB to 10 dB in 5 dB steps. Another subset of 200 recordings were selected to be filtered by 16 real RIRs from AIR database. A subset of 200 recordings were also chosen to represent nonlinear distortions in signals by processing them in a same way the UBM data were distorted. The remaining 200 recordings were kept unchanged to represent the clean samples.

Using a Hamming window, recordings were segmented into frames of 30 ms with 10 ms overlap. For each frame of a signal, 12 MFCCs together with the log-energy are calculated along with delta and double-delta coefficients. They are concatenated to form a 39-dimensional feature vector.

6.1.2. Results

To evaluate the proposed approach in identifying degradations in data not observed during the training phase, we used 10-fold cross validation with 10 iterations. For each experiment, we extended the test subset by adding 20 outliers, which contain irrelevant sounds for PD detection randomly selected from the MMPD data set, to show whether the detectors could reject such outliers. Moreover, as an example of combination of degradations in speech signals, 20 good-quality recordings were selected from the MMPD data set, contaminated by noise and reverberation in a similar way we did for the UBM data, and appended them to the test subset to investigate whether both the noise and reverberation detectors could identify these recordings.

Fig. 8 shows the performance of the detectors in terms of AUC, along with 95% confidence intervals, as a function of the number of mixture components in GMMs. We can observe from the results that the degradations in voice signals are effectively identified when GMMs with 1024 mixtures are used. The lower performance for reverberation detection model is mainly due to misdetection of some of the recordings in which noise and reverberation coexist but the noise is more dominant than the reverberation. This can also be explained by considering the analysis of vowels in the presence of different degradations (Poorjam et al., 2017) which shows that MFCCs of the reverberant signals are, on average, positioned closer to the MFCCs of the clean signals, while noise and distortion (clipping) shift the MFCCs farther away from the position of clean MFCCs.

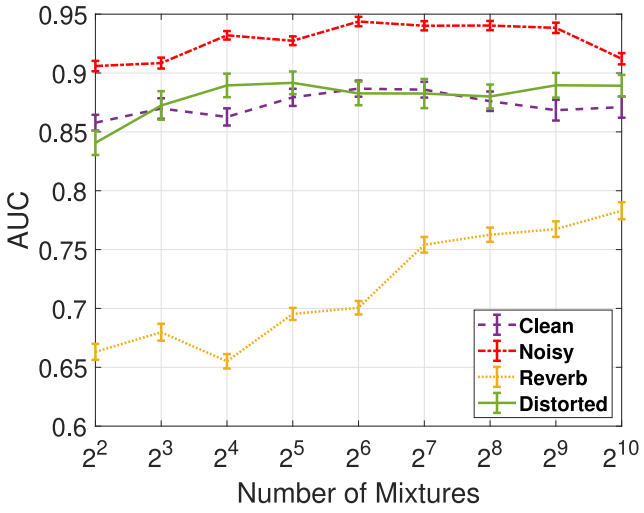


Fig. 8. The performance of the proposed recording-level degradation detection in terms of AUC, along with 95% confidence intervals, as a function of number of mixture components.

6.2. Frame-level quality control

While many types of degradation, such as reverberation and nonlinear distortions, typically influence the entire recording, additive noise and some kinds of nonlinear distortion such as clipping can have a short-term impact on a signal. Moreover, the test protocol can be violated for a short period of time in a remotely collected voice signal. In recording-level degradation detection, we assumed that the majority segments of a voice signal are influenced by some types of degradation. Likewise, if a voice sample is an outlier, the majority segments of the signal are assumed to contain irrelevant information for PD detection. Even though beneficial in providing a global information about the quality of a signal, it does not say whether a degraded or an outlier signal still contains useful segments to be considered for PD detection. Identifying these segments facilitates making the most use of the available data.

In this paper, we consider additive noise and hard clipping as examples of short-term degradations in a signal, and develop a framework which splits a voice signal into variable duration segments in an unsupervised manner by fitting an infinite hidden Markov model (iHMM) to the frames of the recordings in the MFCC domain. Then, the degraded segments and those that are associated with the protocol adherence or violation are identified by applying a multinomial naive Bayes classifier.

A HMM represents a probability distribution over sequences of observations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$ by invoking a Markov chain of hidden state variables $s_{1:T} = (s_1, \ldots, s_t, \ldots, s_T)$ where each s_t is in one of the K possible states (Rabiner, 1989). The likelihood of the observation $\boldsymbol{x}_t$ is modeled with a distribution of K mixture components as:

$$p(\boldsymbol{x}_t | s_{t-1} = i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_{i,k} p(\boldsymbol{x}_t | \boldsymbol{\theta}_k), \tag{5}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$ are the time-independent emission parameters, $\pi_{ij} = p(s_t = j | s_{t-1} = i)$, $(i, j = 1, 2, \dots, K)$, is the transition matrix of $K \times K$. We consider a HMM for clustering the frames of the signals in terms of different acoustic events. The prediction of the number of states required to cover all events such that we do not encounter unobserved events in the future is challenging. Moreover, it is reasonable to assume that as we observe more data, different types of protocol violations and acoustic events will appear and thus the inherent number of states will have to adapt accordingly. Here, we propose to use an infinite HMM to relax the assumption of a fixed K in (5), defined as:

$$\begin{aligned} \boldsymbol{\beta} &\sim \text{GEM}(\gamma) \\ \boldsymbol{\pi}_k &\sim \text{DP}(\alpha, \boldsymbol{\beta}) \qquad (k = 1, 2, \dots, \infty) \\ \boldsymbol{\theta}_k &\sim H \qquad (k = 1, 2, \dots, \infty) \\ s_0 &= 1 \\ s_t | s_{t-1} &\sim \boldsymbol{\pi}_{s_{t-1}} \qquad (t = 1, 2, \dots, T) \\ \boldsymbol{x}_t | s_t &\sim f(\boldsymbol{\theta}_{s_t}) \qquad (t = 1, 2, \dots, T). \end{aligned} \tag{6}$$

where $\boldsymbol{\pi}_k \sim \text{DP}(\alpha, \boldsymbol{\beta})$ are drawn from a Dirichlet process (DP) with a local concentration parameter $\alpha > 0$, $\boldsymbol{\beta}$ is the stick-breaking representation for DPs which is drawn from Griffiths–Engen–McCloskey (GEM) distribution with a global concentration parameter $\gamma > 0$ (Sethuraman, 1994), each $\boldsymbol{\theta}_k$ is a sample drawn independently from the global base distribution over the component parameters of the HMM H, and f is the observation model for each state. The iHMM can possibly have countably infinite number of hidden states. Using the direct assignment Gibbs sampler, which marginalizes out the infinitely many transition parameters, we infer the posterior over the sequence of hidden states $\boldsymbol{\pi}$ and emission parameters $\boldsymbol{\Theta}$. In each iteration of the Gibbs sampling, we first re-sample the hidden states and then the base distribution parameters. For more details about the inference, we refer to Poorjam et al. (2019b). The whole sequence $s_{1:T}$ is sampled for M burn-in iterations followed by N post burn-in iterations. The convergence is verified by inspecting the joint data log-likelihood. Then, the posterior of s is empirically estimated from the samples after convergence. Since s is categorical, its posterior is a histogram of frequencies of state value $k \in \{1, \dots, K\}$ observed for a state indicator s_t in the sampling iterations after burn-in.

Considering an iHMM as a clustering algorithm, segments of the voice recordings with similar characteristics are clustered together under the same state indicator values. To identify the segments that are sufficiently reliable for detecting PD voice symptoms, those that need enhancement before being used for PD detection, and those which do not contain relevant information for PD detection, we propose to use the multinomial naive Bayes classifier to map the state indicators $s_{1:T}$ to the labels $y_{1:T} = (y_1, \dots, y_t, \dots, y_T)$, where $y_t = 1$ if $\boldsymbol{x}_t$ is clean and adheres to the protocol, $y_t = 2$ if it complies with the protocol but is degraded by additive noise, $y_t = 3$ if it is degraded by distortion, or $y_t = 4$ if it violates the protocol and does not contain relevant information for PD detection. Assuming that the samples in different classes have different multinomial distributions, we train the multinomial naive Bayes classifier using the posterior probabilities of the state indicators $s_{1:T}$ of the training data along with the corresponding class labels $y_{1:T}$. The feature vector for the tth observation $\boldsymbol{\rho}_t = (\rho_{t,1}, \dots, \rho_{t,K})$ is a histogram, with $\rho_{t,k}$ being the number of times state k is observed. The likelihood of the histogram of a new observation $\tilde{\boldsymbol{\rho}}$ is defined as:

$$P(\tilde{\boldsymbol{\rho}} | y_{1:T}, \tilde{y}, \boldsymbol{\rho}_{1:T}) = \frac{(\sum_{k=1}^{K} \rho_{t,k})!}{\prod_{k=1}^{K} \rho_{t,k}!} \prod_{k=1}^{K} p_{k,\tilde{y}}^{\rho_{t,k}}, \tag{7}$$

where $p_{k,\tilde{y}}$ is the probability of the kth attribute being in class $\tilde{y} \in \{1, 2, 3, 4\}$ trained using the labeled training data. Using the Bayes rule and the prior class probability $P(\tilde{y})$, the class label for a new test observation is predicted as:

$$\hat{y} = \underset{y \in \{1,2,3,4\}}{\arg\max} \left(\log P(\tilde{y} = y) + \sum_{k=1}^{K} \rho_{t,k} \log(p_{k,y}) \right). \tag{8}$$

Table 2
The confusion matrix of the proposed frame-level quality control method. Results are in the form of mean ± STD.

			Predicted class			
			Adherence	Degraded		Violation
				Noisy	Clipped	
Actual class	Adherence		91% ± 1%	2% ± 0%	4% ± 1%	2% ± 0%
	Degraded	Noisy	16% ± 2%	77% ± 2%	5% ± 1%	2% ± 1%
		Clipped	14% ± 2%	2% ± 1%	82% ± 2%	2% ± 1%
	Violation		8% ± 2%	1% ± 1%	1% ± 0%	90% ± 2%

6.2.1. Experimental setup

To evaluate the performance of the proposed method, a subset of 150 good-quality recordings (representing equally PD patients and healthy controls of both genders) has been selected from the MMPD data set. The quality of this subset is evaluated by manually inspecting the recordings so that no ambient noise, reverberation or distortion is perceived in the signals and that they comply with the test protocol. From this subset, 50 recordings were selected and 60% of each signal were degraded by adding noise. We used babble, office, restaurant, street and wind noises, under different SNR conditions ranging from −5 dB to 15 dB in steps of 2.5 dB. Another 50 recordings were distorted by hard clipping at different clipping levels ranging from 0.2 to 0.7 in 0.05 steps. The remaining 50 recordings of this subset were considered as clean samples that adhere to the test protocol. In addition, 20 recordings from the MMPD data set containing several short- and long-term protocol violations were selected and added to the subset.

Using a Hamming window, recordings are segmented into frames of 30 ms with 10 ms overlap. For each frame of a signal, 12 MFCCs along with the log energy are calculated. The features of every five consecutive frames are averaged to smooth out the impact of articulation (Poorjam et al., 2018a), and to prevent capturing very small changes in signal characteristics, which could result in producing many uninterpretable states. Thus, each observation represents an averaged MFCCs of ≈100 ms of a signal. For the iHMM, we use the conjugate normal-gamma prior over the Gaussian state parameters, set the hyper-parameters $\alpha{=}\gamma = 10$, and run the inference for 200 iterations consisting of 20 burn-in iterations.

6.2.2. Results

The top plot in Fig. 9 shows a segment of 10 s selected from the data set. The noisy segments that need enhancement and the segments of the signal that adhere to the test protocol are hand-labeled and shaded in pink and green, respectively. Fitting the iHMM to the data (i.e. MFCCs of 17,000 frames of 100 ms), 51 different states were discovered in this particular subset. The middle plot in Fig. 9 illustrates the generated states in different colors. To evaluate the performance of the proposed approach for data not observed during the training phase (i.e. out of sample), we used 10-fold CV and repeated the procedure 10 times. The results, presented in Table 2, indicate that the proposed method can automatically identify short-term degradations and protocol violations in pathological voices with a 0.1 s resolution and high accuracy.

6.3. Integrating quality control and enhancement algorithms

The proposed quality control approaches can be integrated with the enhancement algorithms for cleaning-up the remotely collected signals before they are being processed by a PD detection system. In this section, we evaluate how this integration can lead to improvement in PD detection accuracy.

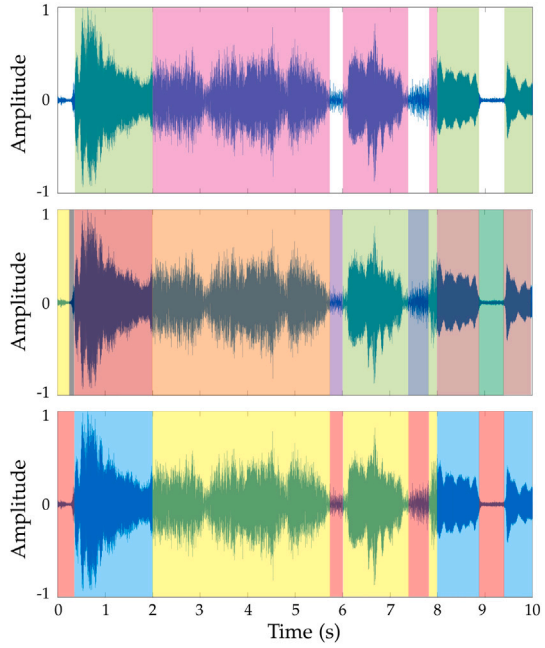


Fig. 9. Illustrative results of applying the proposed frame-level quality control method to a 10-second segment of the voice recordings selected from the data set. In the top plot, the green shaded and pink shaded areas represent the segments of the signal which are hand-labeled as adhering to the protocol and those degraded by the background noise, respectively. The middle plot shows the states, generated by the iHMM, in different colors. The bottom plot illustrates the result of applying a trained classifier to the state indicators to predict which segments adhere to the protocol (shaded in blue), which ones violate the protocol (shaded in red), and which ones are noisy (shaded in yellow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6.3.1. Recording-level

The recording-level algorithm can be used in many different ways to provide information about the presence and type of degradation in a signal for an automatic clean-up process. For example, one possible scenario could be to convert the parallel detectors to a multi-class classifier by calculating the maximum *a posteriori* probability for a new observation. Then, the enhancement algorithm for which the observation has the highest degradation class probability will be applied. Nevertheless, the advantage of the proposed method over the multi-class classification-based techniques is that it can be considered as a multi-label, multi-class classifier that has the capability to detect outlier recordings and those degraded by a new type of degradation. Thus, alternative approach could be to exploit the detectors to activate or bypass a set of enhancement blocks connected in series (e.g. noise reduction followed by dereverberation) or in parallel. This scenario not only allows enhancement of a signal degraded by more than one degradation, but also prevents processing of outliers or recordings degraded by a degradation type for which there is no effective enhancement algorithm. However, since there is no ground truth health status label for the outlier recordings, it is not possible to evaluate the performance of the PD detection system in the presence of outliers. For this reason, we consider a simple scenario in which the test subset only contains clean, noisy, clipped, and reverberant recordings. Since there is no outlier in the test samples, the problem is simplified to a multi-class classification task. For this experiment, we used the same 160 test recordings we used for the enhancement experiments. From this subset, 40 recordings were randomly selected and corrupted by restaurant, office and street noises under different SNR conditions ranging from −5 dB to 7 dB in 4 dB steps. Another 40 randomly chosen recordings were filtered by 16 real RIRs from AIR database. The next 40 randomly selected recordings were clipped with different clipping levels ranging from 0.2 to 0.8 in 0.2 steps. The last 40 recordings were used, without adding degradation, as the clean signals. The DNNSE-N, DNNSE-R, and ASPADE algorithms have been used for noise reduction, dereverberation, and declipping, respectively. Due to the randomness involved in this experiment, it was repeated 100 times.

Table 3
Evaluation of the impact of applying the proposed recording-level quality control in combination with enhancement algorithms on the PD detection performance. Results are in the form of mean AUC ± 95% confidence interval.

Scenarios	AUC
No enhancement	0.86 ± 0.007
Enhancement based on random label assignment	0.87 ± 0.004
Enhancement using DNNSE-NR	0.88 ± 0.004
Enhancement based on predicted labels	0.91 ± 0.004
Enhancement based on ground truth labels	0.91 ± 0.003
Clean signals	0.95

Table 3 shows the PD detection performance in terms of AUC (along with 95% confidence intervals) for five different scenarios: (1) when no enhancement algorithm is applied to the recordings; (2) when the recordings, regardless of the presence and type of degradation, were assigned random labels (clean, noisy, clipped, or reverberant) and processed accordingly; (3) when all recordings, regardless of the presence and type of degradation, were processed by the DNNSE-NR algorithm; (4) when recordings were enhanced by the enhancement algorithm selected based on the estimated degradation labels; and (5) when the degraded recordings were enhanced based on the ground truth degradation labels. The second scenario shows the impact of using the degradation detection with a chance-level of performance. In the third scenario, we ignore the information provided by the degradation detection system and process all the recordings by the DNNSE-NR algorithm. The fifth scenario shows the upper bound for the performance of the PD detection when the degraded signals were enhanced with proper enhancement algorithms. Comparing the first and fifth rows of the table clearly shows that the enhancement could significantly improve the performance of the PD detection system. Similar results in the fourth and fifth rows should not be surprising since the utterance-level degradation detection correctly identified the degradation types 90% of the time. We can conclude from the results that applying appropriate enhancement algorithms to the degraded signals leads to an improvement in PD detection performance, and the level of improvement is related to the accuracy of the degradation detection system.

6.3.2. Frame-level

In the next experiment, we investigate how the proposed frame-level quality control can improve the performance of PD detection. To this aim, we selected 80 random recordings, equally from both classes and genders, and randomly added babble, restaurant, office and street noises at different SNRs ranging from −5 dB to 10 dB in 5 dB steps. It should be noted that, for making a signal noisy, instead of adding a noise to the entire signal, we randomly corrupted 60% frames of the signal. The remaining 80 recordings were clipped at different levels ranging from 0.2 to 0.8 in 0.05 steps. It is worth mentioning that the percentage of the frames of a clipped signal that are affected by clipping is related to the severity of clipping. In both degradation cases, we used the indices of the frame affected by the degradation to produce the ground truth frame-level degradation labels. The enhancement algorithms used for noise reduction in this experiment are the Kalman-CB and the DNNSE-N. The ASPADE is used for declipping.

In Table 4, we compare the PD detection performance for the following scenarios: (1) when no enhancement algorithm is applied to the recordings, (2) when the entire signals are enhanced by either of the enhancement algorithms, (3) when those segments of the signals identified as degraded are enhanced with the corresponding enhancement algorithm, and (4) when the segments of the signals are enhanced with the corresponding algorithms based on the ground truth labels.

Table 4
Evaluation of the impact of applying the proposed frame-level quality control on the PD detection performance, along with 95% confidence intervals.

Scenarios	Algorithms	AUC
No enhancement	–	0.89 ± 0.003
Enhancement of the entire signal	DNNSE-N	0.89 ± 0.004
	Kalman-CB	0.90 ± 0.003
	ASPADE	0.91 ± 0.003
Enhancement based on predicted labels	ASPADE and DNNSE-N	0.93 ± 0.002
	ASPADE and Kalman-CB	0.94 ± 0.001
Enhancement based on ground truth labels	ASPADE and DNNSE-N	0.93 ± 0.002
	ASPADE and Kalman-CB	0.94 ± 0.001
Clean Signals	–	0.95

It is to be noted that since we used two different noise reduction algorithms, the results of the third and the fourth scenarios are reported separately when either of these algorithms was used. For example, the term "*ASPADE and Kalman-CB*" means the Kalman-CB is used for denoising the frames identified as noisy and the ASPADE is used for declipping the frames identified as clipped. For the last two scenarios, we dropped the features of the frames identified/labeled as protocol violation. Due to the randomness involved in this experiment, we repeated the experiment 100 times. The results, reported in the form of mean AUCs ± 95% confidence intervals, show the effectiveness of integrating the proposed frame-level quality control and the enhancement algorithm in dealing with short-term degradation and protocol violations in recordings. Moreover, we can observe that the Kalman-CB algorithm outperforms the DNNSE-N algorithm.

7. Conclusion

Additive noise, reverberation and nonlinear distortion are three types of degradation typically encountered during remote voice analysis which cause an acoustic mismatch between training and operating conditions. In this paper, we investigated the impact of these degradations on the performance of a PD detection system. Then, given that the specific degradation is known, we explored the effectiveness of a variety of the state-of-the-art enhancement algorithms in reducing this mismatch and, consequently, in improving the PD detection performance. We showed how applying appropriate enhancement algorithms can effectively improve the PD detection accuracy. To inform the choice of enhancement method, we proposed two quality control techniques operating at recording- and frame-level. The recording-level approach provides information about the presence and type of degradation in voice signals. The frame-level algorithm, on the other hand, identifies the short-term degradations and protocol violations in voice recordings. Experimental results showed the effectiveness of the quality control approaches in choosing appropriate signal enhancement algorithms which resulted in improvement in the PD detection accuracy in mismatched acoustic conditions. Even though we performed our study on sustained vowels and using a specific PD detection algorithm in which speech signals were parametrized by PLP coefficients and GMM-UBM was used for scoring and classification, we expect that similar trends will hold for running speech, PD detection systems with different parametrization methods and classifiers, and regression-based methods. For example, the PD detection systems that are based on the clinically interpretable features such as pitch, jitter, shimmer, formants, and articulation rate, can benefit from the proposed quality control and enhancement method as the estimation of these features is highly influenced by the degradation in a signal (Lee, 2012). Moreover, the PD detection systems that are based on cepstral coefficients such as MFCCs can also benefit from the proposed method as we have demonstrated in Poorjam et al. (2017) that the amount of change in the distribution of the cepstral coefficients, due to degradation, is correlated with the level of degradation in a signal. However, more research is needed to support this hypothesis.

This study has important implications that extend well beyond the PD detection system. It can be considered as a step towards the design of robust speech-based applications capable of operating in a variety of acoustic environments. For example, since the proposed quality control approaches are not limited to specific speech types, they can be used as a pre-processing step for many end-to-end speech-based systems, such as automatic speech recognition, to make them more robust against different acoustic conditions. They might also be utilized to automatically control the quality of recordings in large-scale speech data sets. Moreover, these approaches have the potential to be used for other sensor modalities to identify short- and long-term degradations and abnormalities which can help to choose an adequate action.

CRediT authorship contribution statement

Amir Hossein Poorjam: Literature review, Designed algorithms, Experiments, Labeled data, Writing - original draft. **Mathew Shaji Kavalekalam:** Literature review, Experiments, Writing - original draft. **Liming Shi:** Experiments, Writing - original draft. **Jordan P. Raykov:** Algorithm design, Writing - review & editing. **Jesper Rindom Jensen:** Algorithm design, Writing - review & editing. **Max A. Little:** Algorithm design, Experimental structure and results, Writing - review & editing. **Mads Græsbøll Christensen:** Supervised the project, Algorithm and the methodology, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was funded by Independent Research Fund Denmark: DFF 4184-00056.

References

Alam, J., Kenny, P., Bhattacharya, G., Kockmann, M., 2017. Speaker verification under adverse conditions using i-vector adaptation and neural networks. In: Interspeech. pp. 3732–3736.
Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65 (4), 943.
Badawy, R., Raykov, Y.P., Evers, L.J.W., Bloem, B.R., Faber, M.J., Zhan, A., Claes, K., Little, M.A., 2018. Automated quality control for sensor based symptom measurement performed outside the lab. Sensors 18 (4).
Bot, B.M., Suver, C., Neto, E.C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E.R., Friend, S.H., Trister, A.D., 2016. The mpower study, Parkinson disease mobile data collected using ResearchKit. Sci. Data 3 (160011).
Brabenec, L., Mekyska, J., Galaz, Z., Rektorova, I., 2017. Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation. J. Neural Transm. 124 (3), 303–334.
Castellano, P., Sradharan, S., Cole, D., 1996. Speaker recognition in reverberant enclosures. In: IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. pp. 117–120.
Deller, J.R., Hansen, J.H.L., Proakis, J.G., 2000. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York.
Eaton, J., Naylor, P.A., 2013. Detection of clipping in coded speech signals. In: 21st European Signal Processing Conference. pp. 1–5.
Eliasova, I., Mekyska, J., Kostalova, M., Marecek, R., Smekal, Z., Rektorova, I., 2013. Acoustic evaluation of short-term effects of repetitive transcranial magnetic stimulation on motor aspects of speech in Parkinson's disease. J. Neural Transm. 120 (4), 597–605.
Erkelens, J.S., Hendriks, R.C., Heusdens, R., Jensen, J., 2007. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. IEEE Trans. Audio Speech Lang. Process. 15 (6), 1741–1752.
Evers, L.J., Krijthe, J.H., Meinders, M.J., Bloem, B.R., Heskes, T.M., 2019. Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS. Mov. Disord. 34 (10), 1480–1487.

Fakhry, M., Poorjam, A.H., Christensen, M.G., 2018. Speech enhancement by classification of noisy signals decomposed using NMF and Wiener filtering. In: 26th European Signal Processing Conference.

Fan, J., Han, F., Liu, H., 2014. Challenges of big data analysis. Natl. Sci. Rev. 1 (2), 293–314.

Gerkmann, T., Hendriks, R.C., 2012. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. 20 (4), 1383–1393.

Gil, D., Johnson, M., 2009. Diagnosing Parkinson by using artificial neural networks and support vector machines. Glob. J. Comput. Sci. Technol. 63–71.

Goh, Z., Tan, K.C., Tan, B.T.G., 1999. Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model. IEEE Trans. Speech Audio Process. 7 (5), 510–524.

Gong, Y., 1995. Speech recognition in noisy environments: A survey. Speech Commun. 16 (3), 261–291.

Habets, E.A., 2006. Room impulse response generator. Tech. Univ. Eindh. Tech. Rep. 2 (2.4), 1.

Habets, E., 2007. Single- and Multi-Microphone Speech Dereverberation Using Spectral Enhancement (Ph.D. thesis). Technische Universiteit Eindhoven.

Habets, E.A., Gannot, S., Cohen, I., Sommen, P.C., 2008. Joint dereverberation and residual echo suppression of speech signals in noisy environments. IEEE Trans. Audio Speech Lang. Process. 16 (8), 1433–1451.

Han, K., Wang, Y., Wang, D., Woods, W.S., Merks, I., Zhang, T., 2015. Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (6), 982–992.

Hansen, J.H.L., Kumar, A., Angkititrakul, P., 2014. Environment mismatch compensation using average eigenspace-based methods for robust speech recognition. Int. J. Speech Technol. 17 (4), 353–364.

He, Q., Bao, F., Bao, C., 2017. Multiplicative update of auto-regressive gains for codebook-based speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (3), 457–468.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87 (4), 1738–1752.

Hu, Y., Loizou, P.C., 2007. Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. 49 (7–8), 588–601.

Huang, Y., Benesty, J., 2003. A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans. Signal Process. 51 (1), 11–24.

Ishihara, L.S., Cheesbrough, A., Brayne, C., Schrag, A., 2007. Estimated life expectancy of Parkinson's patients compared with the UK population. J. Neurol Neurosurg. Psychiatry 78, 1304–1309.

Janssen, A., Veldhuis, R., Vries, L., 1986. Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. IEEE Trans. Acoust. Speech Signal Process. 34 (2), 317–330.

Jeub, M., Schafer, M., Esch, T., Vary, P., 2010. Model-based dereverberation preserving binaural cues. IEEE Trans. Audio Speech Lang. Process. 18 (7), 1732–1745.

Jeub, M., Schafer, M., Vary, P., 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In: International Conference on Digital Signal Processing. pp. 1–5.

Jukić, A., Doclo, S., 2014. Speech dereverberation using weighted prediction error with Laplacian model of the desired signal. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5172–5176.

Jukić, A., van Waterschoot, T., Gerkmann, T., Doclo, S., 2015. Multi-channel linear prediction-based speech dereverberation with sparse priors. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (9), 1509–1520.

Kameoka, H., Nakatani, T., Yoshioka, T., 2009. Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 45–48.

Kavalekalam, M.S., Nielsen, J.K., Boldt, J.B., Christensen, M.G., 2019. Model-based speech enhancement for intelligibility improvement in binaural hearing aids. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (1), 99–113.

Kim, C., Stern, R.M., 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In: Ninth Annual Conference of the International Speech Communication Association.

Kitić, S., Bertin, N., Gribonval, R., 2015. Sparsity and cosparsity for audio declipping: a flexible non-convex approach. In: International Conference on Latent Variable Analysis and Signal Separation. Springer, pp. 243–250.

Kodrasi, I., Doclo, S., 2016. Joint dereverberation and noise reduction based on acoustic multi-channel equalization. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (4), 680–693.

Kowalski, M., Siedenburg, K., Dörfler, M., 2013. Social sparsity! neighborhood systems enrich structured shrinkage operators. IEEE Trans. Signal Process. 61 (10), 2498–2511.

Lee, B.S., 2012. Noise Robust Pitch Tracking by Subband Autocorrelation Classification (Ph.D. thesis). Columbia University.

Mammone, R.J., Xiaoyu Zhang, Ramachandran, R.P., 1996. Robust speaker recognition: A feature-based approach. IEEE Signal Process. Mag. 13 (5), 58.

Mekyska, J., Smekal, Z., Galaz, Z., Mzourek, Z., Rektorova, I., Faundez-Zanuy, M., López-de Ipiña, K., 2016. Perceptual features as markers of parkinson's disease: the issue of clinical interpretability. In: Recent Advances in Nonlinear Speech Processing. pp. 83–91.

Ming, J., 2004. Universal compensation–an approach to noisy speech recognition assuming no knowledge of noise. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1. IEEE, pp. I–961.

Ming, J., Hazen, T.J., Glass, J.R., Reynolds, D.A., 2007. Robust speaker recognition in noisy conditions. IEEE Trans. Audio Speech Lang. Process. 15 (5), 1711–1723.

Mohammadiha, N., Smaragdis, P., Leijon, A., 2013. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans. Audio Speech Lang. Process. 21 (10), 2140–2151.

Moro-Velázquez, L., Gómez-García, J.A., Godino-Llorente, J.I., Villalba, J., Orozco-Arroyave, J.R., Dehak, N., 2018. Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease. Appl. Soft Comput. 62, 649–666.

Murdoch, B.E., 1998. Dysarthria: A Physiological Approach to Assessment and Treatment. Nelson Thornes.

Narayana, M., Kopparapu, S.K., 2009. Effect of noise-in-speech on mfcc parameters. In: Proceedings of the 9th WSEAS International Conference on Signal, Speech and Image Processing, and 9th WSEAS International Conference on Multimedia, Internet & Video Technologies. World Scientific and Engineering Academy and Society (WSEAS), pp. 39–43.

Nercessian, S., Torres-Carrasquillo, P., Martinez-Montes, G., 2016. Approaches for language identification in mismatched environments. In: IEEE Spoken Language Technology Workshop. pp. 335–340.

Nielsen, J.K., Jensen, T.L., Jensen, J.R., Christensen, M.G., Jensen, S.H., 2017. Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient. Signal Process. 135, 188–197.

Orozco-Arroyave, J.R., Arias-Londoño, J.D., Vargas-Bonilla, J.F., Nöth, E., 2013. Perceptual analysis of speech signals from people with Parkinson's disease. Natural Artif. Models Comput. Biol. - Lect. Notes Comput. Sci. 7930 (1), 201–211.

Orozco-Arroyave, J.R., Hönig, F., Arias-Londoño, J.D., Vargas-Bonilla, J.F., Daqrouq, K., Skodda, S., Rusz, J., Nöth, E., 2016. Automatic detection of Parkinson's disease in running speech spoken in three different languages. J. Acoust. Soc. Am. 139, 481.

Pan, Y., Waibel, A., 2000. The effects of room acoustics on mfcc speech parameter. In: Sixth International Conference on Spoken Language Processing.

Poorjam, A.H., Jensen, J.R., Little, M.A., Christensen, M.G., 2017. Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis. In: Interspeech. pp. 289–293.

Poorjam, A.H., Little, M.A., Jensen, J.R., Christensen, M.G., 2018a. A supervised approach to global signal-to-noise ratio estimation for whispered and pathological voices. In: IEEE International Conference on Acoustics, Speech and Signal Processing.

Poorjam, A.H., Little, M.A., Jensen, J.R., Christensen, M.G., 2018b. A parametric approach for classification of distortions in pathological voices. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 286–290.

Poorjam, A.H., Little, M.A., Jensen, J.R., Christensen, M.G., 2019a. Quality control in remote speech data collection. IEEE J. Sel. Top. Sign. Proces. 13 (2).

Poorjam, A.H., Raykov, Y.P., Badawy, R., Jensen, J.R., Christensen, M.G., Little, M.A., 2019b. Quality control of voice recordings in remote Parkinson's disease monitoring using the infinite hidden Markov model. In: IEEE International Conference on Acoustics, Speech and Signal Processing.

Poorjam, A.H., Saeidi, R., Kinnunen, T., Hautamäki, V., 2016. Incorporating uncertainty as a quality measure in i-vector based language recognition. In: Speaker and Language Recognition Workshop. Bilbao, Spain. pp. 74–80.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Ramaker, C., Marinus, J., Stiggelbout, A.M., Van Hilten, B.J., 2002. Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. Mov. Disord.: Off. J. Mov. Disord. Soc. 17 (5), 867–876.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. 10 (1), 19–41.

Reynolds, D., Rose, R., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3 (1), 72–83.

Rusz, J., Hlavnička, J., Tykalová, T., Novotný, M., Dušek, P., Šonka, K., Ružička, E., 2018. Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease. IEEE Trans. Neural Syst. Rehabil. Eng. 26 (8), 1495–1507.

Santos, J.F., Falk, T.H., 2018. Speech dereverberation with context-aware recurrent neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (7), 1236–1246.

Schroeder, M., Atal, B., 1985. Code-excited linear prediction(CELP): High-quality speech at very low bit rates. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 10. pp. 937–940.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statist. Sinica 4, 639–650.

Siedenburg, K., Kowalski, M., Dörfler, M., 2014. Audio declipping with social sparsity. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1577–1581.

Singh, N., Pillay, V., Choonara, Y.E., 2007. Advances in the treatment of Parkinson's disease. Prog. Neurobiol. 81 (1), 29–44.

Srinivasan, S., Samuelsson, J., Kleijn, W.B., 2006. Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 14 (1), 163–176.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4214–4217.

Tsanas, A., Little, M.A., McSharry, P.E., Ramig, L.O., 2010. Accurate telemonitoring of parkinsons disease progression by noninvasive speech tests. IEEE Trans. Biomed. Eng. 57 (4), 884–893.

Tsanas, A., Little, M.A., McSharry, P.E., Ramig, L.O., 2011. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. J. R. Soc. Interface 8 (59), 842–855.

Tsanas, A., Little, M.A., McSharry, P.E., Spielman, J., Ramig, L.O., 2012. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans. Biomed. Eng. 59, 1264–1271.

Varga, A., Steeneken, H.J., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.

Vasquez-Correa, J., Arias-Vergara, T., Orozco-Arroyave, J.R., Vargas-Bonilla, J.F., Arias-Londono, J.D., Nöth, E., 2015. Automatic detection of parkinson's disease from continuous speech recorded in real-world conditions. In: Interspeech. pp. 3–7.

Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P., 2007. First stereo audio source separation evaluation campaign: data, algorithms and results. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (Eds.), Independent Component Analysis and Signal Separation. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 552–559.

Vorländer, M., 2007. Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality. Springer Science & Business Media.

Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (10), 1702–1726.

Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., Kellermann, W., 2012. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. 29 (6), 114–126.

Záviška, P., Mokrý, O., Rajmic, P., 2018. S-spade done right: Detailed study of the sparse audio declipper algorithms. arXiv preprint arXiv:1809.09847.

Záviška, P., Rajmic, P., Ozerov, A., Rencker, L., 2020. A survey and an extensive evaluation of popular audio declipping methods. arXiv preprint arXiv:2007.07663.

Zhan, A., Little, M.A., Harris, D.A., Abiola, S.O., Dorsey, E.R., Saria, S., Terzis, A., 2016. High frequency remote monitoring of Parkinson's disease via smartphone: platform overview and medication response detection. pp. 1–12, arXiv preprint arXiv:1601.00960.

Amir Hossein Poorjam (S'18) received the M.Sc. degree in electrical engineering from Katholieke Universiteit Leuven, Leuven, Belgium, in 2014. In 2016, he joined the Audio Analysis Lab, Aalborg University, Aalborg, Denmark, as a Ph.D. Fellow working on "Speech Pre-processing for Parkinson's Disease Diagnosis." In 2015, he visited the Speech and Image Processing Unit, University of Eastern Finland, Joensuu, Finland, where he was granted the CIMO scholarship. He visited the Mathematics Group, Aston University, Birmingham, U.K., in 2018. His main research interests include machine learning, speech processing, and pathological voice analysis.

Mathew Shaji Kavalekalam was born in Thrissur, India, in 1989. He received the B.Tech. degree in Electronics and Communications Engineering from Amrita University, Coimbatore, India, and the M.Sc. degree in Communications Engineering from RWTH Aachen University, Aachen, Germany, in 2011 and 2014, respectively. In 2019, he received the Ph.D. degree from Aalborg University, Aalborg, Denmark. His research interests include speech enhancement for hearing aid applications.

Liming Shi (S'14) was born in Henan, China, in August 1989. He received the M.Eng. degree in information and communication engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2015. In 2019, he received the Ph.D. degree from Aalborg University, Aalborg, Denmark, where he is currently a postdoctoral researcher. His research interests include audio analysis, adaptive filtering, and Bayesian approaches for signal processing.

Yordan P. Raykov received the B.Sc. in mathematics from University of Leicester, U.K. and the Ph.D. in machine learning from Aston University, Birmingham, U.K. in 2013 and 2016 respectively. During his Ph.D., he developed scalable inference algorithms for some of the most common Bayesian nonparametric models. He then joined the R&D team in ARM Cambridge to work on the development of novel embedded device for occupancy estimation. Following ARM, he did a pharma funded postdoc with Aston University and Radboud unversity medical center where he developed probabilistic models for unsupervised and semi-supervised learning, with focus on health monitoring applications. From 2018, Yordan is an assistant professor in the Mathematics Department in Aston where he studies scalable nonparametric graphical models and hybrid models for latent structure discovery across multiple domains including: health monitoring, medical imaging, proteomics and finance.

Jesper Rindom Jensen (S'09–M'12) was born in Ringkøbing, Denmark in August 1984. He received the M.Sc. degree *cum laude* in 2009 from Aalborg University in Denmark. In 2012, he received the Ph.D. degree from Aalborg University. Currently, he is a Associate Professor at the Department of Architecture, Design and Media Technology (CREATE) at Aalborg University in Denmark. Before this, he was with the Department of Electronic Systems at Aalborg University, and in addition to this he has been Visiting Researcher at the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, at the Friedrich-Alexander Universität Erlangen-Nürnberg in Erlangen, Germany, and at the University of Surrey, UK. At Aalborg University, he is also a member of the Audio Analysis Lab since its foundation in 2012. His research interests include signal processing theory and methods for, e.g., robot and drone audition, and microphone arrays. Examples of more specific research interests within this scope are enhancement, separation, localization, tracking, parametric analysis, and modeling. He has published more than 80 papers on these topics in top-tier, peer-reviewed conference proceedings and journals. Moreover, he is the co-author of two books, namely, "Speech Enhancement – A Signal Subspace Perspective" and "Signal Enhancement with Variable Span Linear Filters". Dr. Jensen has received a highly competitive postdoc grant from the Danish Independent Research Council, has been selected as an AAU Talent at Aalborg University (awarded to young research talents), and has received several travel grants from private foundations. Furthermore, he is an Associate Editor of the EURASIP Journal on Audio, Speech, and Music Processing, is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee, is an Affiliate Member of the IEEE Signal Processing Theory and Methods Technical Committee, and is a Member of the IEEE.

Max A. Little began his career by writing software, signal processing algorithms, and music for video games, followed by a D.Phil. in mathematics at the University of Oxford, Oxford, U.K. After postdoctoral positions in Oxford and co-founding a web-based image search business, he received a Wellcome Trust Fellowship at the Massachusetts Institute of Technology, Cambridge, MA, USA, to follow up on his doctoral research work in biomedical signal processing, where he was selected as a TED Fellow. He is currently an Associate Professor at the University of Birmingham, Birmingham, U.K.

Mads Græsbøll Christensen (S'00–M'05–SM'11) received the M.Sc. and Ph.D. degrees from Aalborg University (AAU), Aalborg, Denmark, in 2002 and 2005, respectively. He is currently with the Department of Architecture, Design and Media Technology, AAU, as a Professor in Audio Processing and is the Head and Founder of the Audio Analysis Lab. He was formerly with the Department of Electronic Systems, AAU, and held visiting positions with Philips Research Labs, ENST, University of California, Santa Barbara, and Columbia University. He has authored or coauthored three books and more than 200 papers in peer-reviewed conference proceedings and journals. He has given multiple tutorials

at the European Signal Processing Conference, the IEEE International Conference on Acoustics, Speech, and Signal Processing, and the Annual Conference of the International Speech Communication Association and a keynote talk at the International Workshop on Acoustic Signal Enhancement. His research interests include audio and acoustic signal processing, where he has worked on topics such as microphone arrays, noise reduction, signal modeling, speech analysis, audio classification, and audio coding. Dr. Christensen has received several awards, including Best Paper Awards, the Spar Nord Foundation's Research Prize, a Danish Independent Research Council Young Researcher's Award, the Statoil Prize, the EURASIP Early Career Award, and an IEEE Signal Processing Society Best Paper Award. He is a beneficiary of major grants from the Independent Research Fund Denmark, the Villum Foundation, and Innovation Fund Denmark. He serves as the Editor-in-Chief of the EURASIP Journal on Audio, Speech, and Music Processing and as Senior Area Editor for the IEEE Signal Processing Letters, and he has previously served as an Associate Editor for the IEEE/ACM Transactions on Audio, Speech, and Language Processing and the IEEE Signal Processing Letters. He is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee and a Founding Member of the EURASIP Special Area Team in Acoustic, Speech and Music Signal Processing. He is a member of the EURASIP and the Danish Academy of Technical Sciences.